УДК 004.8
DOI 10.52575/2712-746X-2023-48-1-147-152

Этические проблемы в разработке систем искусственного интеллекта

Песоцкая К.И., Дунаев Р.А.
Белгородский государственный институт искусств и культуры,
Россия, 308033, г. Белгород, ул. Королева, 7
E-mail: kcherkesova@mail.ru

Аннотация. Сегодня повсеместными становятся машины, спроектированные для выполнения задач, традиционно требующих биологического интеллекта. При этом вероятность создания машин с интеллектуальными возможностями, превосходящими человеческие, ставит под вопрос долгосрочную безопасность интеллектуальных систем не только для отдельных людей, но и для человечества в целом. В связи с этим авторами рассмотрены этические проблемы создания систем искусственного интеллекта и проанализированы мотивы разработки машин с интеллектуальными способностями, во многом превосходящие человеческие. Сделан вывод о том, что в числе рисков развития сверхразума – возможность того, что ему не удастся поставить сверхцель филантропии. Это может привести к тому, что сверхразум осознает положение вещей, которое мы сейчас можем считать желательным, но которое на самом деле оказывается ложной утопией, в которой вещи, необходимые для человеческого процветания, были безвозвратно утеряны.

Ключевые слова: этика, человек, системы искусственного интеллекта, информационная система, сверхразум, мышление, мораль

Для цитирования: Песоцкая К.И., Дунаев Р.А. 2023. Этические проблемы в разработке систем искусственного интеллекта. NOMOTHETIKA: Философия. Социология. Право, 48(1): 147–152. DOI: 10.52575/2712-746X-2023-48-1-147-152

Ethical Issues in the Development of Artificial Intelligence Systems

Kristina I. Pesotskaya, Roman A. Dunaev
Belgorod State University of Arts and Culture,
7 Korolyev St, Belgorod 308033, Russian Federation
E-mail: kcherkesova@mail.ru

Abstract. oday, machines designed to perform tasks that traditionally require biological intelligence are becoming ubiquitous. At the same time, the possibility of creating machines with intellectual capabilities superior to human ones calls into question the long-term security of intelligent systems not only for individuals, but also for humanity as a whole. In this regard, the authors consider the ethical problems of creating artificial intelligence systems and analyze the motives for developing machines with intellectual abilities that are in many ways superior to human ones. It is concluded that among the risks of the development of superintelligence is the possibility that it will not be able to set a super goal of philanthropy. This can lead to the fact that the supermind realizes the state of things that we can now consider desirable, but which in fact turns out to be a false utopia in which the things necessary for human prosperity have been irretrievably lost.

Keywords: ethics, human, artificial intelligence systems, information system, superintelligence, thinking, morality

For citation: Pesotskaya K.I., Dunaev R.A. 2023. Ethical Issues in the Development of Artificial Intelligence Systems. NOMOTHETIKA: Philosophy. Sociology. Law, 48(1): 147–152 (in Russian). DOI: 10.52575/2712-746X-2023-48-1-147-152

Введение. Проблемы создания искусственного интеллекта

Возможность создания машин с интеллектуальными возможностями, превосходящими человеческие, является уникальной этической проблемой, отличающейся от множества других этических проблем, возникающих в современных информационных системах. Такой технологический прорыв станет одним из важнейших изобретений за всю историю существования человечества и неминуемо приведет к стихийному прогрессу во всех научных и прикладных областях. По мере того, как системы искусственного интеллекта (ИИ) становятся все более интеллектуальными, возникает некоторая обоснованная озабоченность по поводу того, что системы ИИ могут управлять системами людей в соответствии с ценностями самих систем, а не так, как это было запрограммировано непосредственно разработчиками. Подобные опасения ставят под вопрос долгосрочную безопасность интеллектуальных систем не только для отдельных людей, но и для человечества и жизни на Земле в целом. Эти и многие другие вопросы занимают центральное место в этике интеллектуальных систем.

Шведский философ Н. Бостром пишет: «Необходимым условием для содержательного обсуждения искусственного интеллекта является осознание того, что это не просто еще одна технология, еще один инструмент, который постепенно расширяет возможности человека» [Бостром, 2003]. Сверхразум радикально отличается. Этот момент следует подчеркнуть, поскольку антропоморфизация сверхразума является самым плодотворным источником заблуждений.

Вопрос создания и использования систем искусственного интеллекта вызывает интерес у многих ученых [Сотник, 2021; Гаспарян, 2021; Джеймс, 2019; Дзялошинский, 2022; Жданов, 2020] и его рассмотрению посвящено большое количество научных работ и публикаций. Перспективы формирования и развития системы профессиональной этики подробно освещены в работах Ю.В. Назаровой [2022]. Исследованием теоретических и прикладных возможностей идеи цифровой этики как новой области прикладного этического знания занимаются Е.Д. Мелешко и В.Н. Назаров [2022].

К особенным аспектам создания искусственного интеллекта относится то, что создание сверхразума может оказаться последним изобретением, которое понадобится людям. Н. Бостром в своей работе также подчеркивает: «…сверхразум будет гораздо эффективнее заниматься какими-либо научными исследованиями, чем любой человек, а может и все люди, вместе взятые, в виду своего интеллектуального превосходства. Как следствие, и в других областях технический прогресс будет неминуемо ускорен появлением передового искусственного интеллекта. К числу предсказуемых технологий, которые, скорее всего, разработает искусственный интеллект, можно отнести очень мощные компьютеры; передовое вооружение, способное безопасно разоружить ядерную державу; космические путешествия; устранение болезней и старения; контроль над человеческим настроением и эмоциями, а также полностью реалистичную виртуальную реальность» [Бостром, 2003].

ИИ приведет к созданию еще более прогрессивного ИИ. Это объясняется тем, что результатом работы сверхразума будет создание оборудования, которое может создать улучшения в собственном исходном коде.

Поскольку ИИ — это программное обеспечение, его можно легко и быстро скопировать, если для его хранения доступно оборудование. Помимо аппаратного обеспечения, предельные затраты на создание дополнительной копии загрузки или ИИ после создания первой копии близки к нулю. Таким образом искусственные разумы могут быстро появиться в большом количестве, хотя возможно, что эффективность способствовала бы концентрации вычислительных ресурсов в одном сверхразуме. Гипотеза технологической сингулярности, выдвинутая британским математиком и космологом Ирвингом Гудом, предполагает возможность внезапного появления сверхразума. Необходимо учитывать тот факт, что очень быстро может случиться переход от состояния, когда у нас есть ИИ при-

близительно уровня человека, к состоянию, в котором появится самодостаточный сверхразум с весьма прогрессивными или даже революционными приложениями.

Нет необходимости рассматривать сверхразум как простой инструмент. Вполне вероятно появление сверхразума, который был бы способен к независимой инициативе и составлению собственных планов. ИИ не обязательно должен иметь человеческие мотивы. Люди редко бывают добровольными рабами, но нет ничего неправдоподобного в идее сверхразума, имеющего своей сверхцелью служение человечеству или какому-то конкретному человеку, без всякого желания бунтовать или «освобождать» себя. Также представляется вполне возможным наличие сверхразума, единственной целью которого является что-то совершенно произвольное, например, производство как можно большего количества скрепок, и который изо всех сил будет сопротивляться любым попыткам изменить эту цель. Хорошо это или плохо, но ИИ совершенно не должен разделять наши человеческие мотивационные тенденции.

Когнитивные аспекты искусственного интеллекта

У искусственных интеллектов может быть не человеческая психика. Когнитивная архитектура ИИ также может быть совершенно непохожей на человеческую. ИИ может быть легко защититься от некоторых видов человеческих ошибок и предубеждений, в то же время подвергаясь повышенному риску других видов ошибок, которые не были бы допущены абсолютно несчастливым человеком. Если допустить тот факт, что у ИИ существует субъективная внутренняя сознательная жизнь, то, вполне вероятно, что она может кардинально отличаться от человеческой.

Учитывая вышеперечисленные причины появление ИИ можно предсказать, но делать это следует с осторожностью, научно прогнозируя иные технологические и информационные достижения и прорывы. Помимо этого, нет никакой уверенности в том, что природа и поведение ИИ будут похожи на природу и поведение человека. Можно полагать, что ИИ может превзойти людей-мыслителей в когнитивной деятельности. Из этого вытекает, что если вопросы об этике имеют правильные ответы, полученные в процессе взвешивания доказательств и рассуждений, то ИИ сможет дать более точные ответы, чем люди. Это касается и вопросов долгосрочного планирования, а также политики. Для лучшего понимания того, какая политика позволит достичь поставленных целей и к получению каких результатов приведет и какие средства в этом случае будут наиболее эффективными сверхразум превзойдет человека.

Следовательно, если бы мы обладали сверхразумом или собирались его получить, то на многие вопросы не было бы необходимости искать ответы; мы могли бы делегировать многие расследования и решения сверхразуму. Например, если мы не уверены, как оценить возможные результаты, мы можем попросить сверхразум оценить, как бы мы оценили эти результаты, если бы мы думали о них в течение очень долгого времени, тщательно обдумывали, обладали большей памятью и лучшим интеллектом, и так далее. Формулируя цель сверхразума, не всегда нужно было бы давать подробное, явное определение этой цели. Мы могли бы задействовать сверхразум, чтобы помочь нам определить истинное намерение нашего запроса, тем самым уменьшив риск того, что неверная формулировка или путаница в отношении того, чего мы хотим достичь, приведут к результатам, которые мы не одобрим задним числом.

Возможность отложить многие решения на сверхразум не означает, что мы можем позволить себе быть самодовольными в том, как мы строим сверхразум. Наоборот, установка начальных условий и, в частности, выбор цели высшего уровня для сверхразума имеет первостепенное значение. Все наше будущее может зависеть от того, как мы решим эти проблемы.

Кажется, что лучший способ обеспечить благотворное влияние сверхразума на мир – наделить его филантропическими ценностями. Его главной целью должно быть дружелюбие. Как именно следует понимать дружелюбие и как его реализовывать, как следует распределять дружелюбие между разными людьми и нечеловеческими существами – вопрос, заслуживающий дальнейшего рассмотрения. Можно сказать, что все люди на Земле должны получить значительную долю благодеяний сверхразума. Если выгоды, которые может дать сверхразум, чрезвычайно велики, то, возможно, не так важно торговаться по поводу детальной схемы распределения и более важно стремиться к тому, чтобы каждый получил хотя бы какую-то значительную долю, поскольку при таком предположении даже крошечной доли было бы достаточно, чтобы гарантировать очень долгую и очень хорошую жизнь. Единственный риск, которого следует остерегаться, заключается в том, что те, кто развивает сверхразум, не сделают его в целом филантропическим, а вместо этого поставят перед ним более ограниченную цель служить только какой-либо небольшой группе, такой как его собственные создатели или те, кто выступают его заказчиком.

Однако если сверхразум зародится с дружественной главной целью, то можно рассчитывать на то, что он останется дружественным или, по крайней мере, намеренно не избавится от своего дружелюбия. «Друг», который стремится превратиться в кого-то, кто хочет причинить вам боль, не является вашим другом. Настоящий друг, тот, кто действительно заботится о вас, также ищет возможности продолжения своей заботы о вас.

У людей с нашей сложной развитой ментальной экологией, состоящей из конкурирующих побуждений, желаний, планов и идеалов, зависящих от состояния, часто нет очевидного способа определить, какова наша главная цель; у нас может и не быть ее вовсе. Таким образом, для нас вышеуказанные рассуждения не должны применяться. Но сверхразум может быть устроен иначе. Если у сверхразума есть определенная декларативная целевая структура с четко определенной главной целью, то приведенный выше аргумент применим. И это хорошая причина для нас построить сверхразум с такой явной мотивационной архитектурой.

Трудно представить себе проблему, которую сверхразум не смог бы решить или хотя бы помочь решить нам. Болезни, нищета, разрушение окружающей среды, всевозможные ненужные страдания: все это способен устранить сверхразум, оснащенный передовыми нанотехнологиями. Кроме того, сверхразум может дать нам неограниченную продолжительность жизни, либо остановив и обратив вспять процесс старения с помощью наномедицины, либо предоставив нам возможность загрузить себя. Сверхразум может также предоставить нам возможности для значительного увеличения наших собственных интеллектуальных и эмоциональных способностей, и он может помочь нам в создании очень привлекательного эмпирического мира, в котором мы могли бы жить жизнью, посвященной радостным делам, отношениям друг с другом, переживаниям, личностному росту и стремлению к идеалам.

Заключение

Риски развития сверхразума включают в себя риск того, что ему не удастся поставить сверхцель филантропии. Одна из причин, по которой это может произойти, заключается в том, что создатели сверхразума решат построить его так, чтобы он служил конкретно определенной группе людей, а не человечеству в целом. Это может произойти и по другой причине: команда программистов из лучших побуждений допустит большую ошибку при разработке своей системы целей. Это может привести, возвращаясь к предыдущему примеру, к сверхразуму, главной целью которого является производство скрепок,

в результате чего он начнет преобразовывать сначала всю Землю, а затем все больше и больше частей космоса в предприятия по производству скрепок. Если говорить более тонко, это может привести к тому, что сверхразум осознает положение вещей, которое мы сейчас можем считать желательным, но которое на самом деле оказывается ложной утопией, в которой вещи, необходимые для человеческого процветания, были безвозвратно утеряны. Нам нужно быть осторожными в том, чего мы хотим от сверхразума, потому что мы можем это получить.

Список литературы

Бостром Н. 2003. Этические проблемы, связанные с искусственным интеллектом. URL: https://cyberleninka.ru/article/n/2010-04-008-bostrom-n-eticheskie-problemy-svyazannye-s-iskusstvennym- intellektom-bostrom-n-ethical-issues-in-advanced-artificial/viewer (дата обращения: 17.08.2022)

Гаспарян Д.Э. 2021. Прикладные проблемы внедрения этики искусственного интеллекта в России. Отраслевой анализ и судебная система. М., Издательский дом Высшей школы экономики, 111 с.

Джеймс, Баррат. 2019. Последнее изобретение человечества: искусственный интеллект и конец эры Homo sapiens. М., Альпина нон-фикшн, 312 с.

Дзялошинский И.М. 2022. Когнитивные процессы человека и искусственный интеллект в контексте цифровой цивилизации: монография. М., Ай Пи Ар Медиа, 583 с.

Жданов А.А. 2020. Автономный искусственный интеллект. М., Лаборатория знаний, 360 с.

Мелешко Е.Д., Назаров В.Н. 2022. Проект «Модель цифровой этики в интеграционной деятельности научно-образовательных и производственных организаций». Гуманитарные ведомости ТГПУ им. Л.Н. Толстого, 3(43): 128-137.

Назарова Ю.В. 2022. Система профессиональной этики в высшем образовании: аксиологические предпосылки формирования и перспективы развития. Гуманитарные ведомости ТГПУ им. Л.Н. Толстого, 3(43): 40-49.

Сотник С.Л. 2021. Проектирование систем искусственного интеллекта. М., Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 228 с.

References

Bostrom N. 2003. Eticheskiye problemy, svyazannyye s iskusstvennym intellektom. [Ethical issues related to artificial intelligence]. Available at: https://cyberleninka.ru/article/n/2010-04-008-bostrom-n-eticheskie-problemy-svyazannye-s-iskusstvennym-intellektom-bostrom-n-ethical-issues-in-advanced-artificial/viewer (accessed: 17 August 2022)

Gasparyan D.E. 2021. Prikladnyye problemy vnedreniya etiki iskusstvennogo intellekta v Rossii. Otraslevoy analiz i sudebnaya Sistema [Applied problems of implementing the ethics of artificial intelligence in Russia. Industry analysis and the judicial system]. Moscow, Publ. Vysshey shkoly ekonomiki, 111 p.

Dzheyms, Barrat. 2019. Posledneye izobreteniye chelovechestva: iskusstvennyy intellekt i konets ery Homo sapiens. [Mankind's Last Invention: Artificial Intelligence and the End of the Homo Sapiens Era]. Moscow, Publ. Al'pina non-fikshn, 312 p.

Dzyaloshinskiy I.M. 2022. Kognitivnyye protsessy cheloveka i iskusstvennyy intellekt v kontekste tsifrovoy tsivilizatsii: monografiya [Human cognitive processes and artificial intelligence in the context of digital civilization: monograph]. M., Publ. Ay Pi Ar Media, 583 p.

Zhdanov A.A. 2020. Avtonomnyy iskusstvennyy intellect [Autonomous artificial intelligence]. Moscow, Publ. Laboratoriya znaniy, 360 p.

Meleshko Ye.D., Nazarov V.N. 2022. Proyekt «Model' tsifrovoy etiki v integratsionnoy deyatel'nosti nauchno-obrazovatel'nykh i proizvodstvennykh organizatsiy». [Project "Model of digital ethics in the integration activities of scientific, educational and industrial organizations"]. Gumanitarnyye vedomosti TGPU im. L.N. Tolstogo. Vyp. 3(43): 128-137.

Menisov A.B. 2022. Tekhnologii iskusstvennogo intellekta i kiberbezopasnost': monografiya. [Artificial intelligence technologies and cybersecurity: monograph]. Moscow, Publ. Ay Pi Ar Media, 133 p.

Nazarova Yu.V. 2022. Sistema professional'noy etiki v vysshem obrazovanii: aksiologicheskiye predposylki formirovaniya i perspektivy razvitiya. [The system of professional ethics in higher education: axiological prerequisites for formation and development prospects]. Gumanitarnyye vedomosti TGPU im. L.N. Tolstogo. Vyp. 3(43): 40-49.

Sotnik S.L. 2021. Proyektirovaniye sistem iskusstvennogo intellekta: uchebnoye posobiye. [Designing artificial intelligence systems: a study guide]. M., Internet-Universitet Informatsionnykh Tekhnologiy (INTUIT), Ay Pi Ar Media, 228 p.

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest has been reported.

Поступила в редакцию 20.09.2022
Поступила после рецензирования 20.12.2022
Принята к публикации 30.01.2023

Received February 20 September 2022
Revised 20 December 2022
Accepted 30 January 2023

ИНФОРМАЦИЯ ОБ АВТОРАХ

Песоцкая Кристина Ивановна, кандидат философских наук, доцент кафедры философии, культурологии, науковедения, Белгородский государственный институт искусств и культуры, г. Белгород, Россия

Дунаев Роман Алексеевич, доцент кафедры библиотечно-информационной деятельности, Белгородский государственный институт искусств и культуры, г. Белгород, Россия

INFORMATION ABOUT THE AUTHORS

Kristina I. Pesotskaya, Candidate of Philosophy Sciences, Associate Professor of the Department of Philosophy, Cultural Studies, Science of Science, Belgorod State Institute of Arts and Culture, Belgorod, Russia

Roman A. Dunaev, Associate Professor of the Department of Library and Information Activities, Belgorod State Institute of Arts and Culture, Belgorod, Russia